ELSEVIER

Computer Networks 31 (1999) 669–692

COMPUTER NETWORKS

Enabling QoS adaptation decisions for Internet applications

Saleem N. Bhatti *, Graham Knight

Computer Science Department, University College London, Gower Street, London WC1E 6BT, UK

Received 16 October 1998

Abstract

We present a network model that allows processing of QoS (quality of service) information about media flows to enable applications to make adaptation decisions. Our model is based around a multi-dimensional spatial representation that allows QoS information describing the flow constructions and QoS parameters – flow-states – to interact with a representation of the network QoS. The model produces reports about the compatibility between the flow-states and the network QoS, indicating which flow-states the network can currently support. The simple nature of the reports allows the application to make decisions, *dynamically*, on which flow-state it should use. The model is relatively lightweight and scaleable. We demonstrate the use of the model by simulation of a dynamically adaptive audio tool. Our work is ongoing.

Keywords: QoS (quality of service); QoS summarisation; Adaptive applications; Internet applications

1. Introduction

In a best-effort network such as the Internet, there is no guarantee of the network QoS that a particular application instance might receive. The QoS may fluctuate due to routing effects or traffic interactions in the network [1,2], or the application may be resident on a mobile host [3]. There is currently great interest in making applications adaptable to changes in network QoS. This is of particular importance for real-time media flows or flows that are sensitive to QoS fluctuations. Much of the attention for this work has focused on mechanisms that allow adaptability of the media flow construction by scaling (e.g. Ref. [4]), filtering (e.g. Ref. [5]) or encoding techniques (e.g. [6,7]). An area that has received little attention is how applications can dynamically select the correct flow construction to match the available network QoS. The application must currently rely on the user to set the correct preferences to allow operation in a particular QoS regime. Although application-specific mechanisms exist to allow some automatic adaptation (e.g. elastic buffering to combat jitter in audio tools), we would like to offer a more general model to allow applications to dynamically adjust their flow construction.

In this paper, we start by considering the interactions required between the user, the application and the network in deciding how flows should adapt (Section 2) and consider existing work (Section 3). We then present our model and a short analysis of its function (Section 4) followed by simulation of three scenarios showing an audio tool that uses our model (Section 5). This is followed by a discussion of the key observations and insights we have gained from this work, so far (Section 6). We end with some short concluding remarks (Section 7).

* Corresponding author. Tel.: +44-171-4193249; E-mail: s.bhatti@cs.ucl.ac.uk.

1389-1286/99/$ - see front matter
PII: S0169-7552(98)00294-3

2. The user, the application and the network

To allow adaptation, there must be interaction between the user, the application and the network. The network must supply information about the QoS that is being provided to the flow. The user must specify preferences that govern the behaviour of the application. It is then up to the application to decide how adaptation should take place based on both of these pieces of information. Our area of interest is shown by the dashed box in Fig. 1. We think of the application as having well-defined **modes** that represent operating conditions for the application. Associated with the applications modes are media flow-states, **QoSStates**, that represent the operating conditions for flows. The **QoSSpace** is our network model and the QoSStates conceptually exist within the QoSSpace. The **QoSEngine** maps network state (derived from real QoS parameter measurements for the flow) into the **QoSSpace**. (We do not consider the QoSEngine in detail in this paper, but it does form part of our work.) The QoSSpace issues QoSReports that contain a state compatibility value (SCV) for each flow-state from the application. The application then combines the SCVs with other application-specific information to make an adaptation decision using an application adaptation function (AAF). We say more about QoSStates, QoSSpace, QoSReports and SCVs in Section 4, and show an example of an AAF in Section 5. In Fig. 1, we show only one flow but many are possible. The definition of a flow is application-specific.

Our goal is to allow the application to make adaptation decisions in response to fluctuations in QoS seen by a flow, but the adaptation process should be under the control of the user.

2.1. Interactions between the user and the application

We do not consider in detail the mechanisms for interactions with the user but we discuss the requirements of the user in allowing the application to make adaptation decisions dynamically.

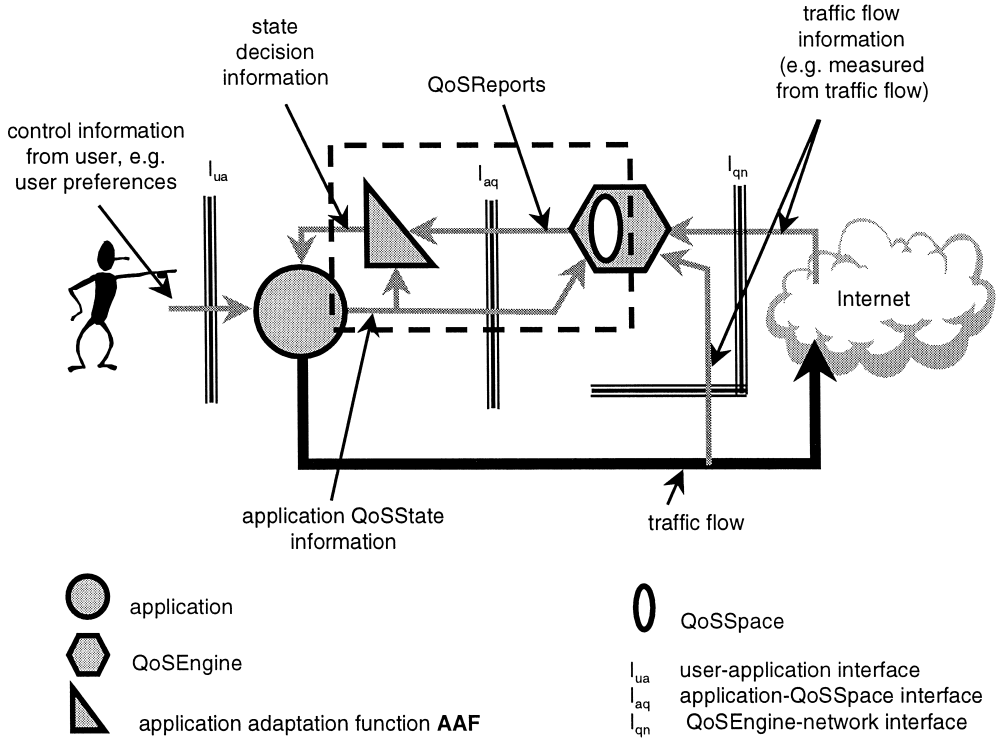

Fig. 1. Interaction between the user, the application and the network.

There may be many instances of an application used by many different users. Each of the users will have their own preferences for the application. Any decision-making process for adaptation must allow the application to apply user preferences to the information about QoS being experienced by the flow. For example, with an audio tool with numerous encoding schemes, one user may specify that the tool ''always uses the best audio quality possible'' while another user may specify that the tool aims to ''retain stability'' in the face of network QoS fluctuations. The application may interpret these preferences as meaning the ''encoding with the highest data rate requirement whenever possible'' and ''do not change flow-state too often'', respectively. So the adaptation decision is not only *application* specific but also *application instance* specific, and controlled by user preferences.

Typically, we would like user preferences to reflect the fact that users may have minimum technical knowledge. User *QoS preferences* may be expressed as an application-specific enumeration such as ''high'', ''medium'' or ''low'' quality. As the application will be making flow-state changes, the user may wish to specify *QoS stability* criterion, e.g. ''do not change flow-state more than once every minute''. Note that the this last criterion is the user's request for the application not to change to a ''better'' quality flow-state too often (i.e. avoid state-flapping), but if the QoS degrades, then we would expect the application to adapt as required.

Such heterogeneity means that the adaptation decision information reflects a *per-instance* view. So the information used by the application to make adaptation decisions should:

- Not unduly constrain the QoS mapping from user requirements to application capabilities.
- Be amenable to the mapping of the user QoS preferences and QoS stability criterion on a per *application instance* basis.

2.2. Interactions between the application and the network

Before an application instance can make decisions about any changes in its behaviour, it needs to know what is ''sensible'' for its flows, i.e. what its current network connectivity can support. This requires some information from the network. Information about the network is typically expressed as values of QoS parameters such as delay, jitter, available capacity, loss, etc. This information may be received in a number of ways:

- Via local mechanisms, e.g. from the communication stack on the host.
- Via application-specific mechanisms, e.g. via proprietary signalling.
- As control messages from the remote receivers of the flow, e.g. using RTP/RTCP [8].
- From network management tools, e.g. using SNMP [9].

If we consider that ''the application knows best'', then we must appreciate that there is likely to be no single ''best'' ubiquitous solution for getting information about the resources within the network. The ''best'' mechanism may depend on the network environment or the application's function or both, but only the application (application designer) is in a position to make that assessment.

The application must also have some way of expressing its functional capability in terms of the construction of its flows, i.e. which flow-states its flows can use. Typically, these will be performance bounds defined in terms of QoS parameter values for the flow-states, e.g. minimum data rate required, maximum delay, etc. Note that as a definition of a flow is application-specific, so is the timescale over which measurements of such performance bounds are measured. This means that our model must be capable of working in whatever timescales are specified by the application.

QoS parameters may include local (end-system) resources, not just network QoS measurements. For example, on mobile systems, measurements of battery power or host load [10,11] may also be used for making adaptation decisions.

So, our network model should be able to process information about QoS parameters, but:

- Should not put any constraints on where that information (measurements) should come from.
- Should be able to accept a wide range of QoS parameters, whether end-system specific or network specific.
- Be able to cope with any timescale over which that information is measured.

2.3. Making adaptation decisions

In Section 2.1, we saw that the user preferences introduce heterogeneity and that the information required to make an adaptation decision may reflect a per-application instance view of QoS. A mechanism is required that can give an indication of the ability of the network to support any of a number of flow-states that an application instance might take. In a distributed application, there may be other application-level signalling involved before adaptation can take place. Also, the application modes may be functions of other application-specific information, so our model can not *make* a decision for the application, but offer *QoS summaries* – **QoSReports** – that represent a view of the relative compatibility of the network QoS and the application's flow capability. However, in general we are not aware of the following application-specific details:

- The nature in which information from a current QoSReport must be evaluated with information from previous QoSReports in order to make flow-state change decisions.
- The nature in which QoSReports must be evaluated with other (application-specific) information in order to make flow-state change decisions.

There may be a statistical or temporal sense in which our QoSReports have meaning to the application with respect to its current mode or flow-states. The application may have synchronisation constraints between its application flows. Such matters can only be assessed by the application. So, we chose to separate the network-related flow-state information from other application-specific information. We chose that the flow-states be expressed as **QoSStates**, flow-state information that is specific only to the QoS requirements and constraints of the individual application traffic flows. For example, an audio application may have several different audio encoding techniques and each would be a separate QoSState for that audio flow. Furthermore, as we do not know how the application will use the QoSReports from our model, we should simply issue QoSReports that are based on an assessment of the *instantaneous* QoS of the network – a snapshot in time. This is also consistent with our requirement for our model to work within the timescale constraints of the application.

Additionally, our model should aim to relieve the application from having to deal with raw QoS parameter measurements [1]. Our aim is to offer the application a method of assessing the need to make flow-state changes, so the granularity at which we report information from our model should be per-flow. The report of the per-flow state information needs to be a suitable summary of the QoS parameter measurements for a flow. In producing the summary, we must be careful not to lose the significance of any single QoS parameter that defines the flow state (e.g. a mean might ''smooth out'' a low or high QoS parameter value).

So, the QoSReports our model offers to the application are per-flow QoS summaries that:

- Should be in a form that can be easily interpreted by a distributed application if required.
- Are derived form QoS parameter values, indicating the relative compatibility between the (instantaneous) network QoS and of an application's QoSStates.

3. QoS assurance and dynamic adaptation

For the provisioning of end-to-end QoS, the user/application is telling the network what is required and asks that the network should adapt/configure itself to comply to the application's requirements, i.e. static adaptability. In enabling dynamic adaptability, the application is constantly (i.e. within application-specific time scales) receiving information about the QoS that the network can offer, and then changing its flow-state to comply with the network capabilities.

3.1. QoS assurance

In Ref. [12], there is an excellent description of the elements of a general QoS architecture for assur-

[1] This does not preclude the use of such measurements for fine-tuning if mechanisms such as elastic buffering are in use, but we are working at a coarser level of flow granularity and over different timescales. Indeed, the value of the QoSParams produced by our QoSEngine back-end may be more suitable for such use than raw values.

ing QoS. With QoS assurance mechanisms, applications have to specify what they want the network to do for them, and the network tries to honour this request. The network itself may provide adaptability mechanisms [13] to cope with fluctuating QoS or heterogeneity due to network architecture [12]. suggests that (among other things) the application should be in control of the following information that passes to the network: QoS service-level (e.g. guaranteed-service, controlled-load service) and QoS management policy (how adaptation should take place when there are QoS fluctuations). These affect the way the application is integrated with the network (the separation of functionality between the network and the application), and how heterogeneity is supported (at the user, application and network access level).

QoS assurance mechanisms are based around the use of resource reservation and QoS re-negotiation to achieve the desired QoS for the application. Mechanisms such as RSVP [14] are designed to provide resource reservations in the Internet. However, RSVP can fail [15], and adaptability mechanisms in the network become impracticable and increasingly difficult to support in ad hoc or mobile network architectures. Additionally, where resource reservation and QoS re-negotiation are used, they should be available end-to-end for maximum utility. In today's Internet there is very little resource reservation support and network components lack the programmability required to enable adaptability in the network. Furthermore, as such mechanisms do become available, they will not be deployed uniformly, and this will further compound the heterogeneity in the Internet. For many applications, such QoS assurance or adaptability mechanisms within the network may only be useful if they are available end-to-end. Where these mechanisms do exist, there is great value in being able to use them, but in the ''standard'' best-effort Internet, other mechanisms are required to address adaptability.

The QoS service-level required may be important to the individual users and may determine the cost of the service, e.g. guaranteed-load is a ''better'' service than controlled-load so may cost more. We argue that the service-level should not be specified by the application. The application should be prepared to be more flexible in its adaptation capability, leaving service-level selection to the user. Three reasons for this are:

1. The service-level may determine the cost of the service and users usually wish to control how much they pay.
2. Network heterogeneity, lack of resource reservation or network element failure may mean that a particular service-level is not available at a given time at a given point in the network.
3. New, additional service-level definitions may be introduced that are more suitable (in terms of functionality or cost) for use in a given situation (e.g. there are two relatively recent descriptions of *adaptive* service-level [16,17]), and differentiated services may be set up that are domain/administration specific [18].

Additionally, if cost-based feedback is available from the network, then cost could be treated as a QoS parameter – although not related to the performance of the flow it acts as a defining constraint in the same way as, say, defining minimum data rate or maximum jitter for a flow. The value of making the cost explicit, as a QoS parameter, is that it highlights the importance of cost as a feedback control mechanism in future services [19].

3.2. Dynamic adaptation

QoS management policy will be subject to user preferences and application-specific behaviour. Applications may find it useful to have a specification of the QoS management policy before the application starts operating. This would certainly be of value to the network for controlled allocation of resources, and makes sense in the context of trying to *assure* end-to-end to QoS. However, in our consideration of dynamic adaptability, the use of the application typically requires interaction with the user in order to determine its adaptation requirements, and these may not be known until after the application is running. In Ref. [12], the QoS management policy: *captures the degree of QoS adaptation (continuous or discrete) that the flow can tolerate and the scaling actions to be taken in the event of violations to the contracted QoS* [20].

We chose to make a separation between what ''*the flow can tolerate*'' and the ''*scaling actions to be taken*''. We argue that the former is a property of

the media and the latter is an application-specific requirement that includes interaction with the user. Flow performance specifications can be used to indicate the flow-states that are possible for a flow and can be determined by the application designer. The action to be taken on fluctuations (''*violations*'') of QoS is a dynamic adaptation decision and cannot be determined by the application designer a priori. It is the difference between the application designer saying, ''I know what is sensible for the application'', and the user saying, ''I know what is sensible for the application to do for me''. Ultimately the application's functional constraints have the final say on which flow-state(s) is (are) functionally *possible*, but this should not dictate *how* the user would like the application to behave, i.e. how adaptation should take place. Ref. [21] points out that the user requirements and the network QoS may change throughout the session and proposes that the user should be given the opportunity to make informed decisions about application adaptation. As an example, consider a remote teaching scenario and the requirements of the audio and video flows. When operating in lecture mode (main part of the teaching session), the conferencing application may tolerate relatively high delay and throughput is asymmetric, but during a question and answer session (at the end of the teaching session), low delay and jitter are required with symmetric throughput for flows.

The application must be able to assess the user preferences and available network QoS in order to make automatic and dynamic adaptation decisions.

4. The QoSSpace

Our concern is that we try and assess the overall relative *compatibility* between the flow QoSStates and the network QoS rather than try to evaluate the *equality* (or otherwise) of the absolute values of the QoS parameters for a flow-state and the measurements taken from the network. We would like our model to indicate *how well* the network QoS matches the requirements for any of an application's flow-states.

4.1. QoSSpace, QoSParams and QoSStates

The **QoSSpace** is a multi-dimensional space in which flows conceptually exist, and into which the network QoS is mapped. Flows are represented by **QoSStates**, and each flow may have a set of QoSStates. The dimensions of the QoSSpace are represented by a set of **QoSParams**. This is depicted in Fig. 2(a), which shows only three QoSParams, but any number of QoSParams are possible.

The QoSParams are variables that are representations of real QoS parameters, such as throughput, delay, jitter, etc. These are chosen to suit the application, i.e. the dimensions of the QoSSpace are application-specific. QoSSpace effectively gives a snapshot in time, and the interval over which measurements are taken will be application-specific. The network QoS is evaluated and mapped into the QoSSpace by some appropriate, application-specific mechanism. We chose to separate the abstraction of the QoSSpace from the mechanisms performing the

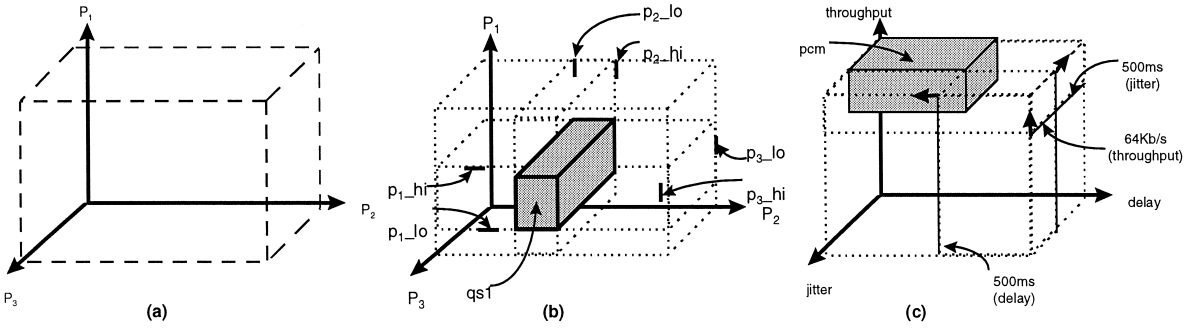

Fig. 2. QoSSpace; (a) example with 3 QoSParams; (b) example QoSState; (c) a single QoSState for an audio tool.

mapping of the network QoS [2]. An example QoSState is shown in Fig. 2(b). This spatial description effectively defines a region in QoSSpace in which the flow can operate. We define a QoSState for a fictitious flow in terms of the QoSParams P_1, P_2 and P_3. We use simple rectilinear boundaries, which we call ***_hi*** and ***_lo***:

$$qs1 = \{\langle P_1, p_1\_lo, p_1\_lo, p_1\_qlo, p_1\_qhi\rangle, \langle P_2, p_2\_lo, p_2\_lo, p_2\_qlo, p_2\_qhi\rangle, \langle P_3, p_3\_lo, p_3\_lo, p_3\_qlo, p_3\_qhi\rangle\}.$$

This statement identifies a QoSState called *qs1*, defined by tuples for the QoSParams P_1, P_2 and P_3. *qs1* consists of a set of tuples, each of which has the structure:

$$\langle id, \_lo, \_hi, \_qlo, \_qhi\rangle$$

where *id* a name identifying the QoSParam, *_lo* low threshold value of the QoSParam, *_hi* high threshold value of the QoSParam, *_qlo* QoSiState low threshold, *_qhi* QoSiState high threshold.

(We defer a description of QoSiStates until Section 4.7.) For any application flow:

- QoSStates for a flow need not be specified with the same number of QoSParams.
- For any QoSParam tuple, either *_hi* or *_lo* threshold may be left undefined, but one must be present.
- QoSStates may overlap.

As an example of a QoSState for an (imaginary) audio tool, consider the definition of the QoSState *pcm* in Fig. 2(c):

$$pcm = \{\langle throughput, 64, -, -, -\rangle, \langle delay, -, 500, -, -\rangle, \langle jitter, -, 500, -, -\rangle\}.$$

This says that the state *pcm* requires a minimum of 64 Kb/s throughput, and can tolerate a maximum of 500 ms delay and a maximum of 500 ms jitter. Other QoSStates may be possible. Note that the model is not concerned with the semantics of the QoSParams, or any relationships between them – this is left for the application to control. Also, different QoSParams may exhibit strong correlation from the QoS parameters that they are derived from (e.g. on some networks, delay and throughput may be related – as delay goes up throughput goes down), and so it may be possible to reduce the number of dimensions. This, again, is an application level issue. The application may choose the complexity of the definition of the QoSStates, as required. However, the QoSSpace must be defined by the set of QoSParams that are the union of all the QoSParams for all the QoSStates for that flow.

4.2. NetQoSState

We now need a mapping of the network QoS into the QoSSpace. We achieve this by using QoSParams to describe the network QoS, **NetQoSState**. The NetQoSState will have the same dimensions as the QoSSpace. QoSParam values are estimates of the current value of the QoS parameter, based on real network measurements of QoS parameters for a flow. A simple interpretation of the NetQoSSpace would be to use the values of the QoSParams. This would translate to a point within the QoSSpace:

$$q = \{p_n\} \quad \forall n = 1, \ldots, N,\ p_n \in P_n,$$

i.e. $\{p_n\}$ are the values of the set of QoSParams $\{P_n\}$ that define the QoSSpace However, delay and noise effects mean that associated with the QoSParams is an *uncertainty* in our QoSParam values. As the network QoS fluctuates we may have different degrees of uncertainty. If the network is in a steady state, we may have a lesser degree of uncertainty than if the network is currently showing fluctuations in the QoS offered. Measures such as standard deviation are normally used to indicate such uncertainty. However such measures only have meaning in a statistical sense when we have some knowledge of the model of our traffic and/or the network. What we are actually after is an indication of the current *variability*, v_p, of the QoSParam. We discuss how we estimate a value for v_p later, but for now we see that in terms of our QoSSpace, we chose the

[2] In our work, the QoSParam values for the network QoS, NetQoSState (see Section 4.2), are generated by the QoSEngine back-end, but we do not describe this in detail here. Indeed, other application-specific mechanisms may be used in place of the QoSEngine back-end.

mapping of the NetQoSState to be expressed as follows:

$$q = \{\langle P_1, p_1\_lo, p_1\_hi\rangle,$$
$$\langle P_2, p_2\_lo, p_2\_hi\rangle, \langle P_3, p_3\_lo, p_3\_hi\rangle\}$$

where the *_lo* and *_hi* thresholds indicate the limits of our estimate of variability of the QoSParam. Note that this is a similar format as our expression for the QoSStates. However, the NetQoSState, must have both a *_lo* and *_hi* component for each QoSParam tuple.

4.3. Assessing compatibility between a QoSState and NetQoSState

In terms of the QoSSpace, a description of the task for our model is relatively straightforward: we need to find when the region defined by the NetQoSState intersects with a region defined by a QoSState. If we can assign a meaningful value to this intersection, we can offer the application a state compatibility value (SCV), a measure of the how well the current network QoS might support a particular QoSState. This compatibility value is a unitless number that is easy to use in other parts of the application.

Note that the definition of the QoSState and that for the NetQoSState suggest that we may be able to treat these state definitions as hyper-volumes. For example, we may chose to use the ratio:

$$\frac{\text{volume of overlap of NetQoSState with QoSState}}{\text{volume of NetQoSState}}$$

to evaluate a SCV for each QoSState. However, we choose not to do this. In definitions of QoSStates, our model allows use of different numbers of QoSParams in defining QoSStates for the same flow, resulting in different shaped volumes. In the evaluation of a volume, relative scaling by multiplication of values of N QoSParams may lead to a distortion when some values are particularly high or particularly low. Indeed, we need to process each QoSParam individually, and then offer some sensible summary to the application. So, we must first consider how we process individual QoSParam values.

4.4. The parameter compatibility value (PCV)

For each QoSParam tuple, we can derive a parameter compatibility value (PCV), that expresses the amount by which the value of a certain QoSParam from the NetQoSState falls within the operating region given by a particular tuple for a given QoSState. So, for a given tuple, T_p, from a QoSState, and the corresponding tuple, T_q, from the NetQoSState, for the same type of QoSParam, P:

$$T_p = \langle P, p_p\_lo, p_p\_hi\rangle,$$
$$T_q = \langle P, p_q\_lo, p_q\_hi\rangle,$$
$$\text{PCV} = \text{PCVF}(T_q, T_p)$$

where PCVF is the parameter compatibility value function. The operation of this function is to assess the following statement:

$$\text{PCVF}(T_q, T_p) = T_q \text{ WITHIN } T_p$$

where WITHIN is an operator that evaluates to a single number in the range [0, 1], for the ratio:

$$\frac{\text{intersection of length of } T_q \text{ and length of } T_p}{\text{length of } T_q}.$$

The description of WITHIN is explained with the help of Fig. 3. This shows the possible scenarios for evaluating WITHIN when T_q and T_p overlap. I is the length of the intersection of T_q with T_p. It is clear that the omission of either p_p_lo or p_p_hi from T_p (form the QoSState) poses no problem.

The PCVF has a simple algorithm:

$$L_q = p_q\_hi, p_q\_lo,$$
$$I = \text{MIN}(p_q\_hi, p_p\_hi) - \text{MAX}(p_q\_lo, p_p\_lo),$$
$$\text{PCV} = \text{MAX}(0, I/L_q).$$

The MIN and MAX functions in line 2 perform their usual operations, except that if either p_p_hi or p_p_lo are not defined, then p_q_hi or p_q_lo are used, respectively, as required. I takes the range $[-\infty, L_f]$. When there is no intersection, I is negative, and we choose that PCV = 0, indicating ''no compatibility'' between QoSState and NetQoSState, while PCV = 1 indicates ''full compatibility''. The final line of the PCVF ensures that PCV is in the range [0, 1].

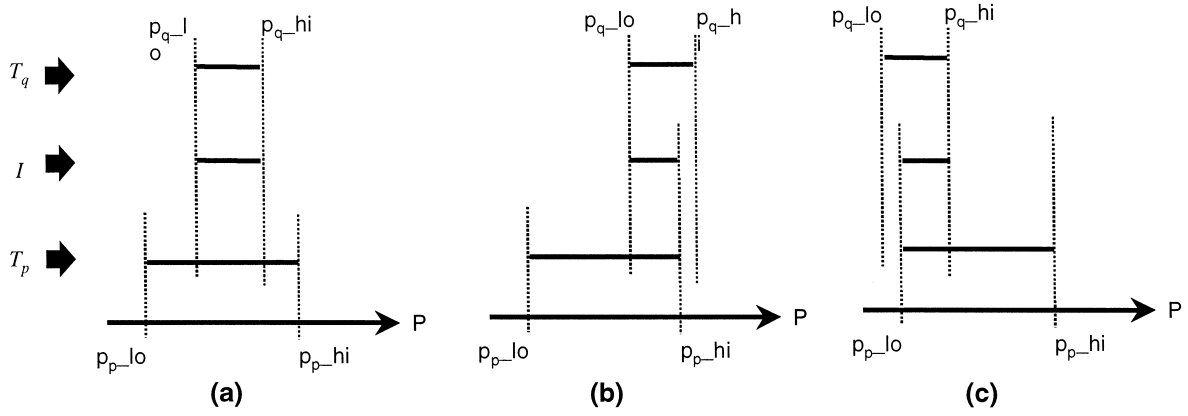

Fig. 3. Scenarios for the evaluation of WITHIN.

This normalised value is a uniform, consistent and scaleable way of representing PCVs. The algorithm for the PCVF is also computationally simple. We see that as the variability of the QoSParam (i.e. the QoS fluctuation) increases, so L_q increases. Unless the fluctuations are completely contained within the region defined by *_lo* and *_hi* tuple thresholds (Fig. 3(a)), as L_q increases, we have decreasing compatibility between the QoSState and NetQoSState with respect to that particular QoSParam (Fig. 3(b) and (c)).

4.5. The variability of QoSParam values

We choose a very simple measure for variability, v_p, to allow us to evaluate p_q_lo and p_q_hi: we chose v_p to be the difference between the current (p_t) and previous (p_{t-1}) QoSParam values:

$$v\_p = \mathrm{ABS}(p_t - p_{t-1}),$$

$$p_q\_lo = p_t - (v\_p/2),$$

$$p_q\_hi = p_t + (v\_p/2).$$

If we were to use p_t and p_{t-1} to produce an *instantaneous* estimate of the standard deviation, σ, of P, we would have

$$\mu = \frac{p_t + p_{t-1}}{2},$$

$$\sigma^2 = \frac{(p_t - \mu)^2 (p_{t-1} - \mu)^2}{2}.$$

Some manipulation gives

$$\sigma = \left| \frac{p_t - p_{t-1}}{2} \right|.$$

So, our measure of variability, v_p, can be seen as a range of $[-\sigma, +\sigma]$ (given by $[p_q\_lo, p_q\_hi]$) about our current value, p_t.

4.6. The state compatibility value (SCV)

The **state compatibility value function (SCVF)** must take the PCVs for all the QoSParams in the QoSState and transform them into a SCV. Any of the usual arithmetic summarisation functions that combine the values (such as a mean), will provide an incorrect PCV due to relative scaling. For example, consider five QoSParams, that give rise to the set of PCVs, $S_A = \{1.0, 1.0, 1.0, 1.0, 0.0\}$. The application may decide that it is likely to take a SCV of 0.8 to consider that a flow-state is useable. This may seem reasonable, but we can see that the mean of S_A is 0.8, yet clearly one of the parameter conditions cannot be supported (hence the PCV of 0.0). The application would make an incorrect decision and this could lead to application and/or network instability.

Looking at it another way, we have seen, in (4.4) that the PCVF evaluates the function WITHIN. For N QoSParams and a QoSState with tuples T_{pn} and NetQoSState with tuples T_{qn} ($\forall n = 1, \ldots, N$), we

Please note a correction for the equation at http://goo.gl/5h9Yo There should be a '+' sign between the two expressions in brackets in the numerator of the right hand side.

base a SCVF algorithm on the following statement from:

if
T_{q1} WITHIN T_{p1} **and**
T_{q2} WITHIN Tp_2 **and** ...
T_{qN} WITHIN T_{pN}
then SCV is HIGH

This reasoning makes linguistic sense. We see that if the QoSParam values (the NetQoSState) *all* fall WITHIN the thresholds defined in the tuples for the corresponding QoSStates, then the degree to which the T_{qn} tuples are WITHIN their corresponding T_{pn} tuples is the degree to which the SCV is HIGH. The key to this assessment is how to evaluate **and** so generate a value for HIGH.

Fuzzy logic provides a suitable interpretation of **and** as the MIN function. So, we can modify our statement to say:

$$\begin{aligned} \text{SCV} = \quad & \text{PCVF}(T_{q1}, T_{p1}) \ \mathbf{and_F} \\ & \text{PCVF}(T_{q2}, T_{p2}) \ \mathbf{and_F} \ \ldots \\ & \text{PCVF}(T_{qN}, T_{pN}) \end{aligned}$$

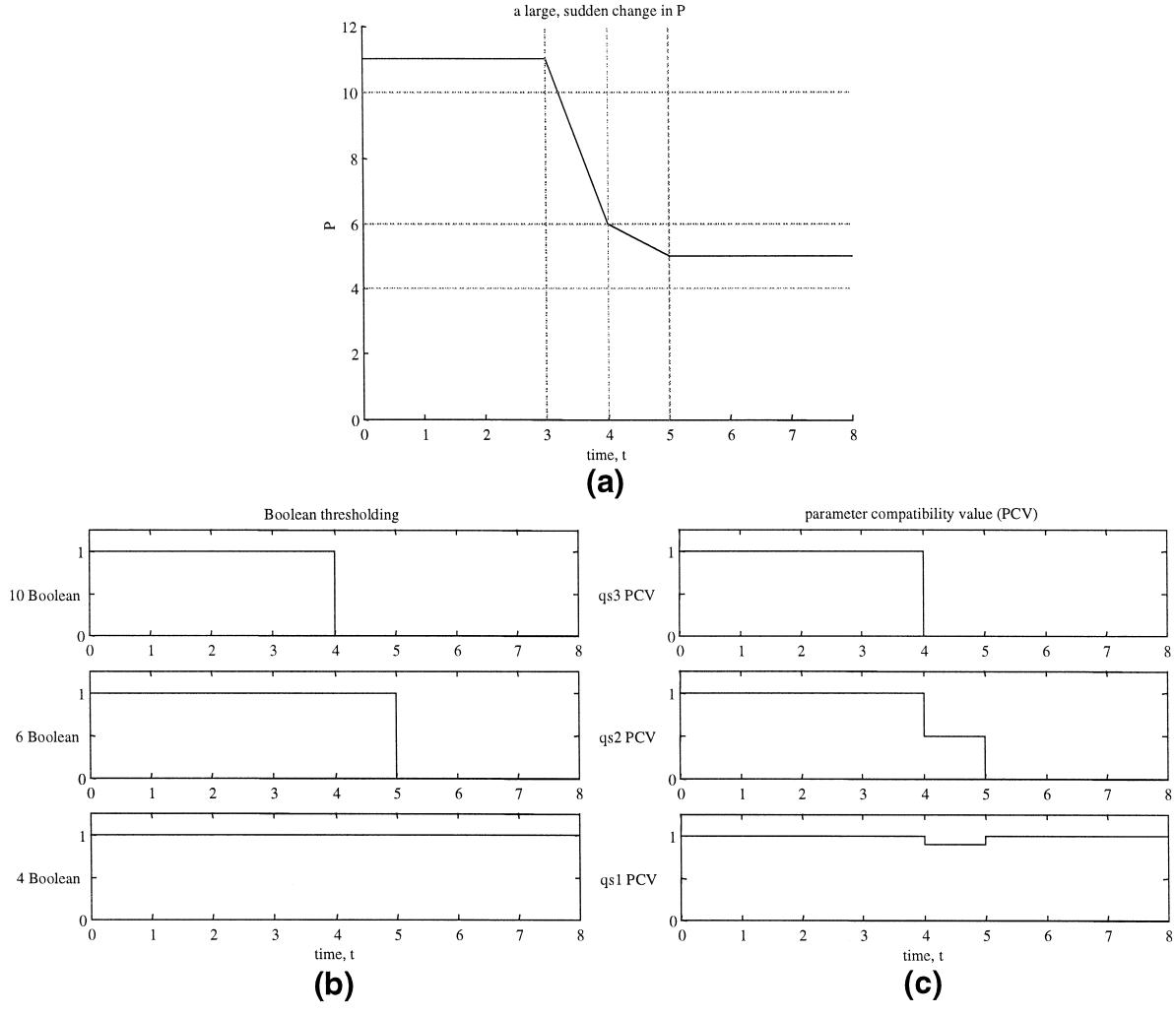

Fig. 4. Comparison of Boolean thresholding and PCVs.

where $\mathbf{and}_{\mathbf{F}}$ is the fuzzy AND operator (MIN). We see from our example for the set of SCVs, S_A, (above) that we would now have the correct behaviour. The use of $\mathbf{and}_{\mathbf{F}}$ (MIN) means that the SCV is in the range [0, 1], and is also computationally simple to evaluate.

Where a QoSState does not have a tuple defined for a particular QoSParam, then for the purposes of the SCVF, this is ignored. (Equivalently, the PCVF can evaluate a PCV = 1.0 for that QoSParam. This is also correct because it means that the QoSState is not dependent on that particular QoSParam, so it always has maximum compatibility with respect to that QoSParam.)

The effect of the using the PCV instead of just simple thresholding (to generate Boolean true/false indications of whether *p_p* is within *p_lo* and/or *p_hi* boundaries) is shown in Fig. 4. We define three simple QoSStates, each with a single tuple:

$$qs3 = \{\langle P,10,-,-,-\rangle\},$$

$$qs2 = \{\langle P,6,-,-,-\rangle\},$$

$$qs1 = \{\langle P,4,-,-,-\rangle\}$$

which are shown marked as horizontal lines at $P = 10$, $P = 6$ and $P = 4$, respectively, in Fig. 4 (a). The area of interest is between times $t = 3$ and $t = 5$. Here we see a sudden drop in P, from a point where it can support all three QoSStates to a point where it can only support *qs1*. In the Boolean thresholds shown in Fig. 4 (b), we see that we have high confidence for *qs2* at $t = 4$ even though we know that the variability of P is high at that point. In contrast, at $t = 4$, the PCV for *qs2* is not high, as shown in Fig. 4(c). At $t = 5$, both the Boolean threshold and the PCV converge, but we see that the use of the PCV may help the application to avoid state-flapping (unnecessarily going into a transient state).

4.7. QoSiStates

Although our model can detect large, sudden changes close to QoSState boundaries, it can not spot slow gradual changes until they cross a boundary (this behaviour arises from our chosen definition of *v_p*). The application may like to have an indication when a QoSParam value in the NetQoSState is nearing a corresponding QoSState tuple *_lo* or *_hi* threshold. This would mean that the application is operating near a QoSState boundary (and so may soon need to change state). We can detect this through the use of **QoS intermediate states** or **QoSiStates**. These are quasi-states that do not represent a flow state as a QoSState does, but are an indication of the proximity of the QoSParam value to the *_lo* or *_hi* threshold of a QoSState. The QoSiState only exists within a QoSState, and is an optional part of the QoSState definition. A QoSiState is also specified by use of a boundary, *_qlo* or *_qhi*. The relationship of the QoSiState to the QoSState with respect to a single QoSParam, P, is depicted in Fig. 5. A QoSiState can only exist if it has a corresponding *_lo* or *_hi* defined in the QoSState.

There are parameter compatibility values, PCV*_hi* and PCV*_lo*, associated with the *_hi* and *lo* QoSiStates, respectively, for each QoSState tuple. These are evaluated in exactly the same way as a PCV for the QoSParam but using the QoSiState tuple, $T_{qq} = \langle P,p\_qhi,p\_hi\rangle$ or $T_{qq} = \langle P,p\_qlo,p\_qlo\rangle$ in place of the QoSState tuple, $T_p = \langle P,p\_lo,p\_hi\rangle$. The *_qhi* and *_qlo* boundaries are specified by the application, and can be left undefined.

The QoSiStates also have a state compatibility value, SCV_I, which is evaluated in a different manner to that of the SCV for the QoSState. Consider a QoSState defined using N QoSParams. If *any* of the N QoSParams tuples, T_{qn}, in the NetQoSState suggest that the corresponding QoSParam value might be WITHIN a QoSiState tuple, T_{qqn}, then we know that the whole QoSState is operating close to one of its boundaries. The application may be cautious and conclude from this that the state may soon not be supported, but this is an application-level

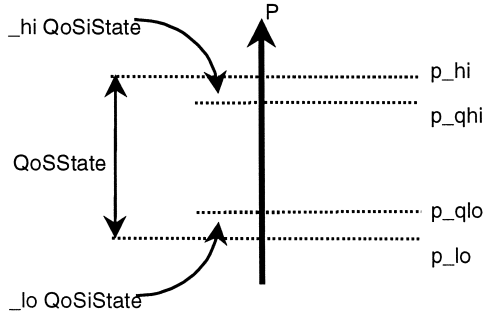

Fig. 5. QoSiStates.

decision. So, our reasoning for evaluating the SCV_I is:

if
T_{q1} WITHIN T_{qq1} **or**
T_{q2} WITHIN T_{qq2} **or** ...
T_{qN} WITHIN T_{qqN}
then SCV_I is HIGH.

In similarity with the evaluation of the SCV, we see that this makes (linguistic) sense. We need an interpretation for **or** that lets us assign a value to SCV_I. Again, fuzzy logic offers us an interpretation of **or** as the MAX function. So, we have

$$\begin{aligned} \text{SCV\_I} = \quad & \text{PCVF}(T_{q1}, T_{qq1}) \ \mathbf{or}_{\mathbf{F}} \\ & \text{PCVF}(T_{q2}, T_{qq2}) \ \mathbf{or}_{\mathbf{F}} \ \ldots \\ & \text{PCVF}(T_{qN}, T_{qqN}) \end{aligned}$$

where $\mathbf{or}_{\mathbf{F}}$ is the fuzzy OR operator. The $\mathbf{or}_{\mathbf{F}}$ operator (MAX) makes sense: it only needs one QoSParam to enter a QoSiState to indicate that the QoSState is operating near a boundary.

In Fig. 6 we demonstrate the use of QoSiStates. We use a *_lo* QoSiState for *qs3* with $p\_qlo = 10.5$, marked as a horizontal line at $P = 10.5$ in Fig. 6(a). We can see in the lower graph of Fig. 6(b) how the value of PCV_*lo* shows that *qs3* is in the region of the QoSiState, ''forewarning'' of a possible state change.

Notice the spike at $t = 50$ in the PCV and PCV_*lo* graph of Fig. 6(b). This may seem to be ''incorrect'' behaviour as we can see that the value of P at that time is within the operating region for *qs1*. However, we only know it is ''correct'' because we can see what happens at $t = 51$. At $t = 50$, all that we can see is that there is a downwards change in the value of P very close to its boundary. We do not know at $t = 50$ what will happen at $t = 51$ so our model assumes that the variability in the value of P will remain high. There is no reason why we can not assume that the value at $t = 51$ will be the *same* as at $t = 50$, but we choose to side with inertia(!) and offer the counter example in Section 4.3 and Fig. 4. We will see how the spike at $t = 50$ can be removed by the application (if required) in Section 5.

5. Dynamic adaptation

In this section, we describe how our model is used to enable dynamically adaptable applications. To demonstrate the use of the QoSSpace, we describe a fictitious audio tool that can adapt its audio flow data rate in response to information about data rate availability for that flow. It does this by changing the audio encoding it uses. We will refer to our example audio tool as ***daat* (dynamically adaptable audio tool)**. *daat* is modelled on information presented in Ref. [22] for an audio tool developed at

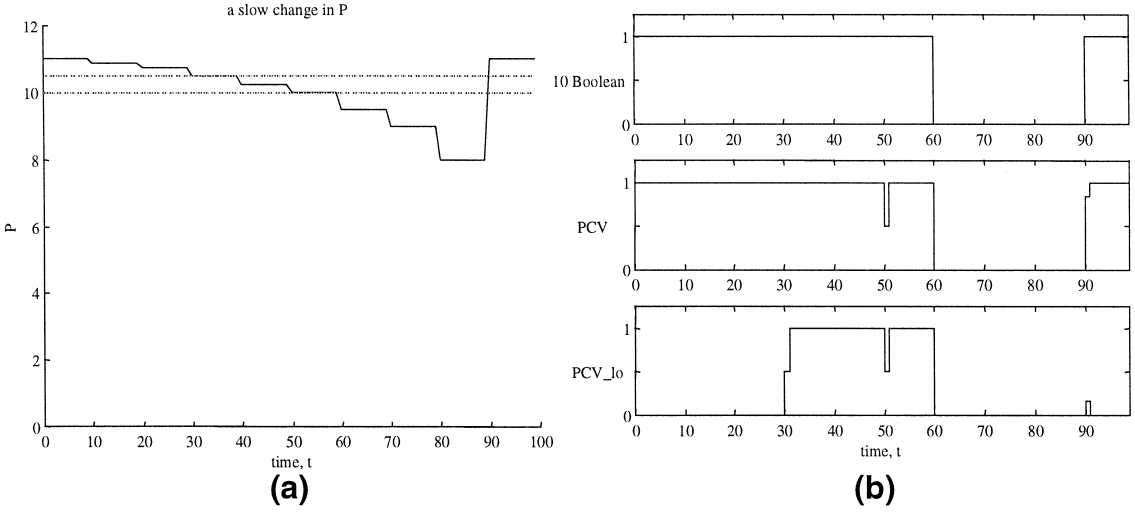

Fig. 6. Using QoSiStates to detect operation close to a QoSState boundary.

Table 1
Flow-state information for daat

Encoding name	Data rate of flow [Kb/s]	Relative CPU cost	Power cost
PCM	64.0	1	0.00001
ADM6	48.0	13	0.00014
ADM4	32.0	11	0.00011
ADM2	16.0	9	0.00009
GSM	13.0	1200	0.01250
LPC	4.8	110	0.00114

UCL [23]. Ref. [6] shows that mixing audio encodings in an audio flow provides usable quality audio streams for Internet wide use. *daat* is capable of the voice encoding schemes shown in Table 1, where the first three columns are taken from [22]. The fourth column is generated from the third column by dividing by 96000, and presents (artificial) power consumption [3] per time unit, for each encoding scheme.

We will assume that an instance of *daat* is running on a mobile host and consider this *daat* instance in isolation. We define two QoSParams, R, the data-rate available to the flow in bits/s, and B, the battery power available on the host. B takes values in the range [0, 1], 1 indicating that the battery on the mobile host is fully charged or that the mobile host is operating on mains power, and 0 indicating that there is no power. We also use the third column of the table to generate some *_lo* thresholds for B by dividing by 2400. We generate *_qlo* thresholds for R by adding 10% to the corresponding *_lo* threshold value. We choose to produce the following QoSStates for *daat* [4]:

$\langle pcm, \langle R, 64.0, -, -, - \rangle \rangle$
$\langle adm6, \langle R, 48.0, -, 52.8, - \rangle, \langle B, 0.0054, -, -, - \rangle \rangle$
$\langle adm4, \langle R, 32.0, -, 35.2, - \rangle, \langle B, 0.0046, -, -, - \rangle \rangle$
$\langle adm2, \langle R, 16.0, -, 17.6, - \rangle, \langle B, 0.0038, -, -, - \rangle \rangle$
$\langle gsm \langle R, 13.0, -, 14.3, - \rangle, \langle B, 0.5, -, -, - \rangle \rangle$
$\langle lpc, \langle R, 4.8, -, -, - \rangle \rangle$

We will consider three scenarios for *daat*:
1. Audio rate adaptation.
2. Audio rate adaptation and power conservation.
3. Helping to prevent state-flapping.

Ref. [22] presents a combined error and rate control mechanism, and we chose only to demonstrate a rate control feature. If we use an additional QoSParam based on packet loss figures we could produce a suitable demonstration of combined error and rate control. We chose to demonstrate only the rate control to simplify our presentation so that the dynamics of our model can be observed.

We first introduce the application adaptation function (AAF) for *daat*. The AAF is based around a threshold value to apply to SCVs, called ***q_compatibility***, and a stability time that governs the rate of QoSState changes to a ''better quality'' QoSState, called ***q_time***. We set $q\_compatibility = 0.8$ and $q\_time = 60$ s. These values are taken as a user policy that says, ''do not change to a better state unless you have 80% compatibility over a one minute period for a better state''. Here, ''better'' is defined by numbering the QoSStates with the values 1–6, with 6 representing the QoSState with the highest data rate requirement (*pcm*) and 1 representing the QoSState with the lowest data rate requirement (*lpc*). We then use the definitions and the AAF shown in pseudo-code in Fig. 7.

In our example, AAF we have chosen to use a very simple mechanism for evaluating the SCVs form the QoSReports. We chose to evaluate mean SCV and SCV_I values for each QoSState over a 60 s time window. These mean values are called the *Q_SCORE* and the *Q_ISCORE* for SCVs and SCV_Is, respectively. The AAF performs the following function:

- If the application has just started ($n < q\_n\_time$), then choose the highest quality (highest numbered) QoSState with the highest *Q_SCORE*.

[3] These artificial values have been generated only for the purposes of our simulation.

[4] We choose to show that the QoSSpace can cope with heterogeneous definitions of QoSStates.

```
if n < q_n_time                                  // just started ..

   q_score_hi = 0,     q_s = 0
   for s = 1 to S                                // higher states override lower ones
        q_s = Q_SCORE(s,n,n)
        if (q_s >= q_score_hi) & (q_s >= q_compatibility)     // use state with highest score
               sdi(n)      = s
               q_score_hi = q_s
               q_epoch     = n
        endif
   endfor

else                              // has been going for some time ..

   if sdi(n-1) > 0                // and was previously in a state
        q_scv = Q_SCORE(sdi(n-1),n,q_n_time)
        q_scv_i = Q_ISCORE(sdi(n-1),n,q_n_time)
   else
        q_scv   = 0
        q_scv_i = 0
   endif

   q_n = n - q_epoch
   if (q_n <= q_n_time) &                                                  // if within epoch and ..
        (q _scv >= q_compatibility) & (q_scv_i < q_compatibility)          // state is still usable ..
        sdi(n) = sdi(n-1)                                                  // stay in present state

   else                           // else check for state change
        for s = 1 to S            // higher states override lower ones
               q_scv = Q_SCORE(s,n,q_n_time)
               q_scv_i = Q_ISCORE(s,n,q_n_time)
               if (q_scv >= q_compatibility) &        // Q_SCORE OK for this state and ..
               (q_scv_i < q_compatibility)            // Q_ISCORE OK for this state so ..
                      sdi(n) = s                      // use this state
               endif
        endfor

        if sdi(n) <> sdi(n-1)                         // state changed so mark new epoch
               q_epoch = n
        endif

   endif
endif
```

Fig. 7. AAF for *daat*: definitions and pseudo-code.

- If the application has been underway for sometime ($n \geq q\_n\_time$):
  - If we are within one minute (q_n_time) of the last QoSState change (q_epoch), do not change QoSState unless the current QoSState is no longer usable.
  - If we are over one minute (q_n_time) since the last QoSState change (q_epoch), check if it is possible to move to another QoSState.

The *Q_SCORE* and *Q_ISCORE* are mean values, and this helps to smooth small disturbances in the SCV and SCV_I respectively, e.g. due to residual noise from the QoS parameter measurements that has passed through the PCVF.

5.1. Audio rate adaptation for daat

Here we use measurements captured from the Internet using ICMP ECHO [ICMPv4] probes [5] that have been processed by the QoSEngine to produce

[5] This produces particularly noisy measurements to demonstrate the robustness of our QoSEngine–QoSSpace system. The values of R that we use were created by processing the raw measurements with our QoSEngine.

values for R. For this scenario, we ignore the QoS tuples defined in terms of B as we assume that the mobile host is operating on mains power. (For these data sets, the values of R are such that the *Q_ISCORE* has no significance on the flow-state change, so we ignore this, too).

We use two data sets, each with probes at 1-s intervals, between the following hosts:

- Darhu–theakston: darhu is a Sun SPARCClassic running SunOS4.1.3 at UCL and theakston is a WinNT4.0 workstation connected to UCL via a single BR-ISDN (Basic Rate ISDN) channel at 64 Kb/s.
- Waffle–tmnserver: waffle is a Sun4 running SunOS4.1.3 at UCL and tmnserver.misa.ch is an IBM server running AIX at IBM Laboratories, Zurich.

In Fig. 8 and Fig. 9: (a) shows the SCV values for the QoSStates. (b) shows the QoSState selected by the AAF. (c) shows the value of R and also the data rate of the audio flow using the AAF.

We see how the audio flow rate is fairly well matched to the available data rate (i.e. the white-space under the line in Fig. 8(a) and Fig. 9(a)), even though R fluctuates considerably. We see from Fig. 9(c) ($600 < \text{time} < 800$), how changes to better quality QoSStates are controlled to the $q\_time = 60$ s time period, but how changes to lower quality QoSStates occur as soon as an inadequate QoS provision is detected. Increasing the value of *q_time* acts to further smooth the data-rate fluctuations, as the *Q_SCORE* is a mean of the SCV values. The effects of using $q\_time = 90$ s, $q\_time = 120$ s and $q\_time = 180$ s for R is shown in Fig. 10. However, increasing *q_time* may make the application sluggish in response to QoS fluctuations. The value of *q_time* could be application-specific, user defined, or its value could also be dynamically adaptable (an issue for further study).

As the value of *q_compatibility* approaches 1.0, this makes a good *Q_SCORE* harder to achieve. This will make it harder for the application to move into (and stay in) better quality states, as shown in Fig. 11. We have found that $q\_compatibility = 0.8$ appears to provide a reasonable threshold, but the exact value chosen may depend on the flow, e.g. video may require a higher *q_compatibility* value than audio.

5.2. Audio rate adaptation and power conservation for daat

We now use artificial values of R and B to demonstrate a (naïve) power conservation policy on the mobile host using *daat*. Here we consider that the mobile host moves from the user's home to the user's place of work. A battery powers the mobile host until reaching work, when mains operation is possible. During battery operation, we wish to conserve power. We use a naïve scheme in which the *_lo* thresholds for each QoSState in the tuples of B mark the lowest battery power charge that is allowed before that audio encoding can be used [6]. The simulation of this scenario is shown in Fig. 12. (For clarity, in the values of R and B, we have not simulated any noise.)

At home our user connects via ISDN (64 Kb/s, $\text{time} < 20$), then by GSM (13 Kb/s, $20 \leq \text{time} < 120$) on the way to work, and finally by Ethernet at work. The user starts with a fully charged battery ($\text{time} = 0$) and is not connected to the mains power until $\text{time} > 200$. We see in Fig. 12(a) and Fig. 12(c) how the *daat* switches from the GSM encoding to the LPC encoding (even though the GSM rate is still achievable) when the battery power, shown in Fig. 12(b), goes down to 0.5 ($\text{time} = 80$). We have used $q\_time = 1$ and $q\_compatibility = 0.8$.

This example shows that the battery power takes precedence in the adaptation policy only when its role becomes significant, i.e. when a QoSState for a boundary defined by QoSParam B is reached. This nature of the QoSState definitions means that we do not need to use relative-scaling or weighting-factors for the PCVFs for individual QoSParams. As long as the QoSState boundaries are defined with appropriate values then the model treats a PCVF threshold as a limit for the QoSState as a whole, and not just a limit for the one QoSParam for which the PCVF generates a low value.

[6] A better scheme might be for the application to perform a calibration when it is first started, by using each encoding scheme to encode a buffer of data. This would measure the rate at which each encoding drains the battery power and then set these values as *_lo* thresholds for a QoSParam that is the *rate of change* in B rather than the absolute value of B itself. We chose our naïve approach in order to demonstrate better the dynamics of the AAF.

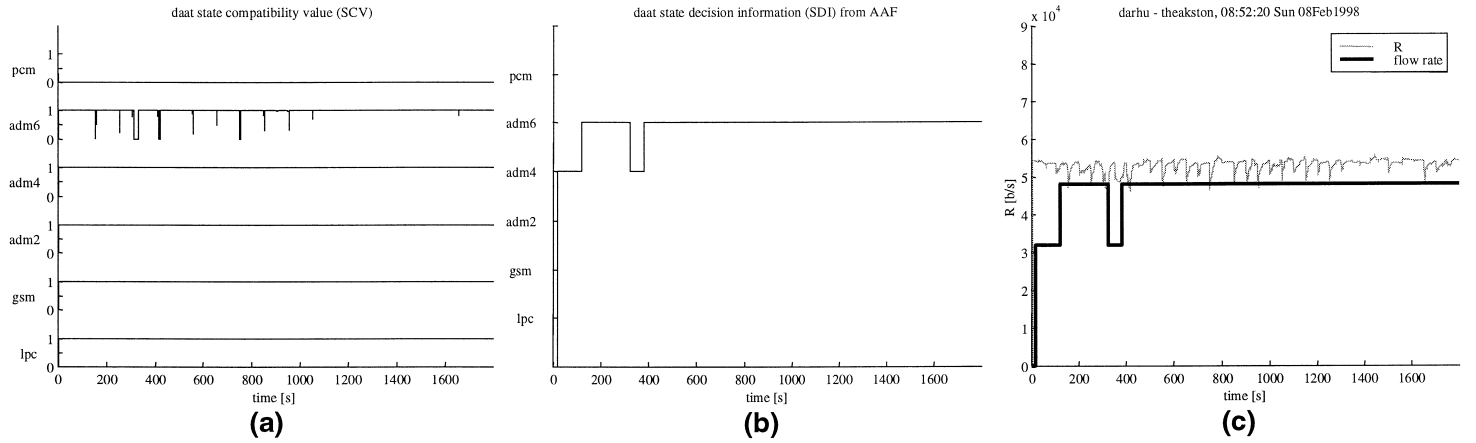

Fig. 8. *daat* with the darhu–theakson data (*q_compatibility* = 0.8, *q_time* = 60 s).

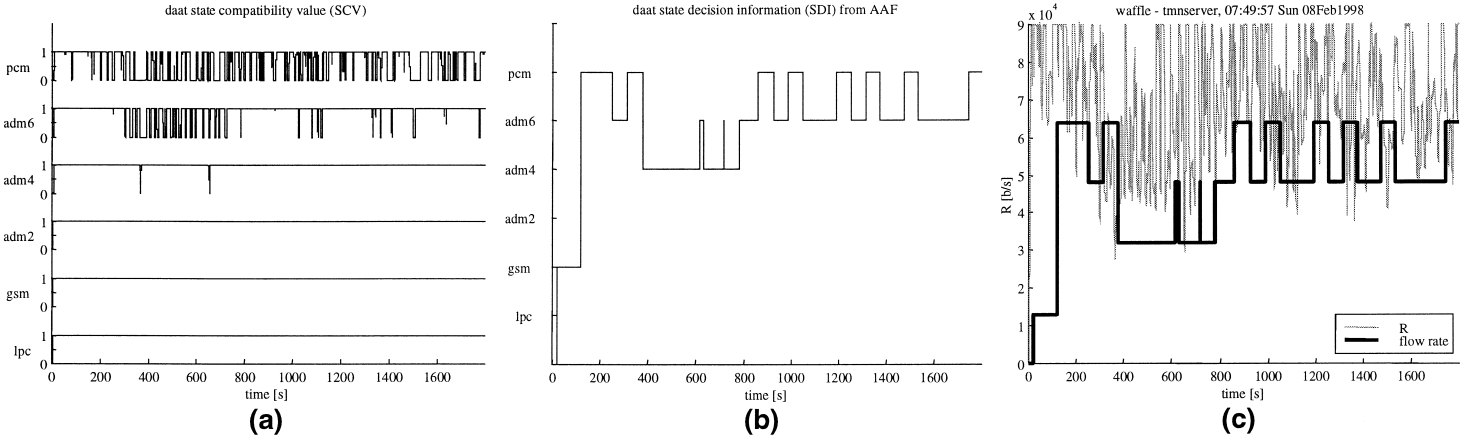

Fig. 9. *daat* with the waffle–tmnserver data (*q_compatibility* = 0.8, *q_time* = 60 s).

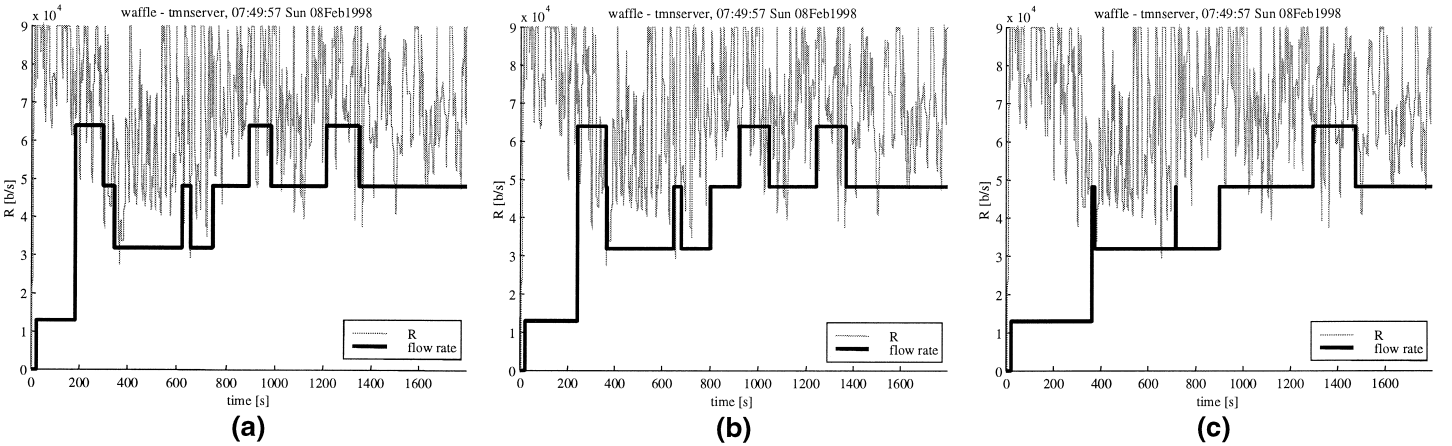

Fig. 10. *daat* with the waffle–tmnserver data (*q_compatibility* = 0.8 and (a) *q_time* = 90 s (b) *q_time* = 120 s (c) *q_time* = 180 s.

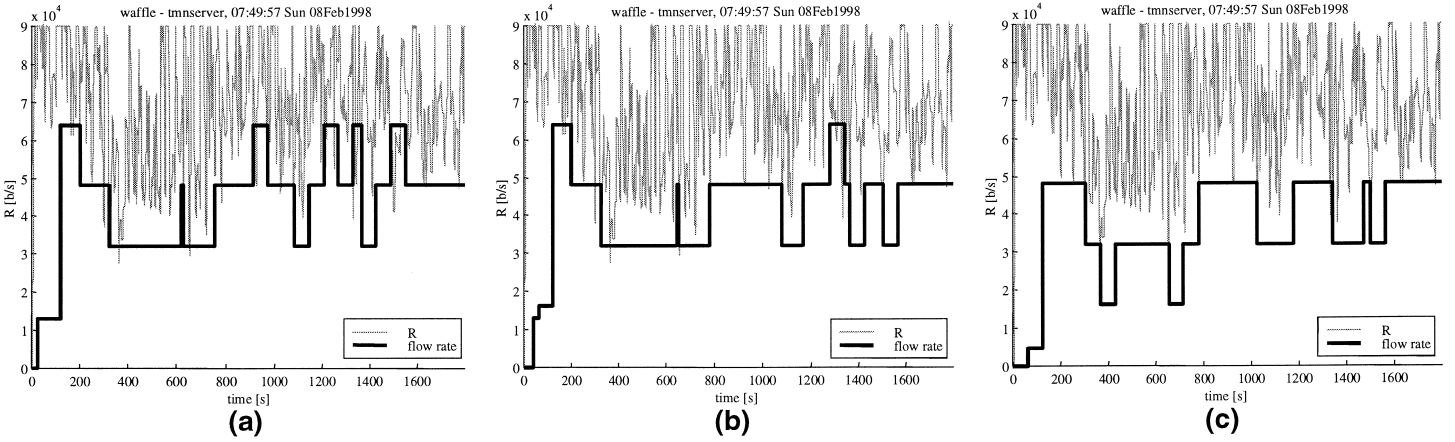

Fig. 11. *daat* with the waffle–tmnserver data (*q_time* = 60 and (a) *q_compatibility* = 0.85 (b) *q_compatibility* = 0.90 (c) *q_compatibility* = 0.95.

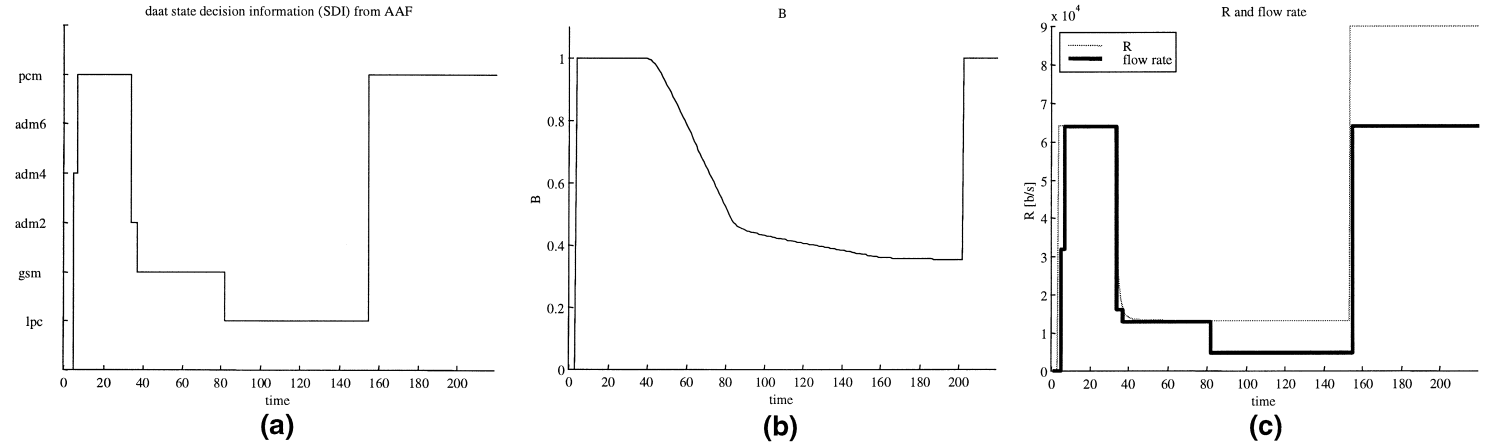

Fig. 12. Artificial rate adaptation and power conservation scenario for daat.

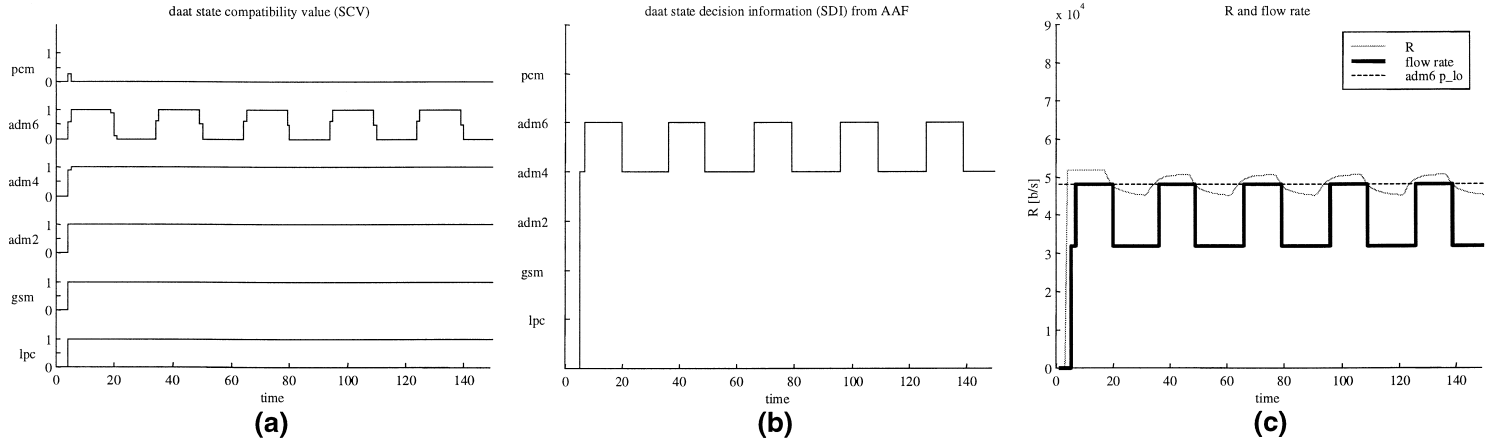

Fig. 13. *daat* with state-flapping.

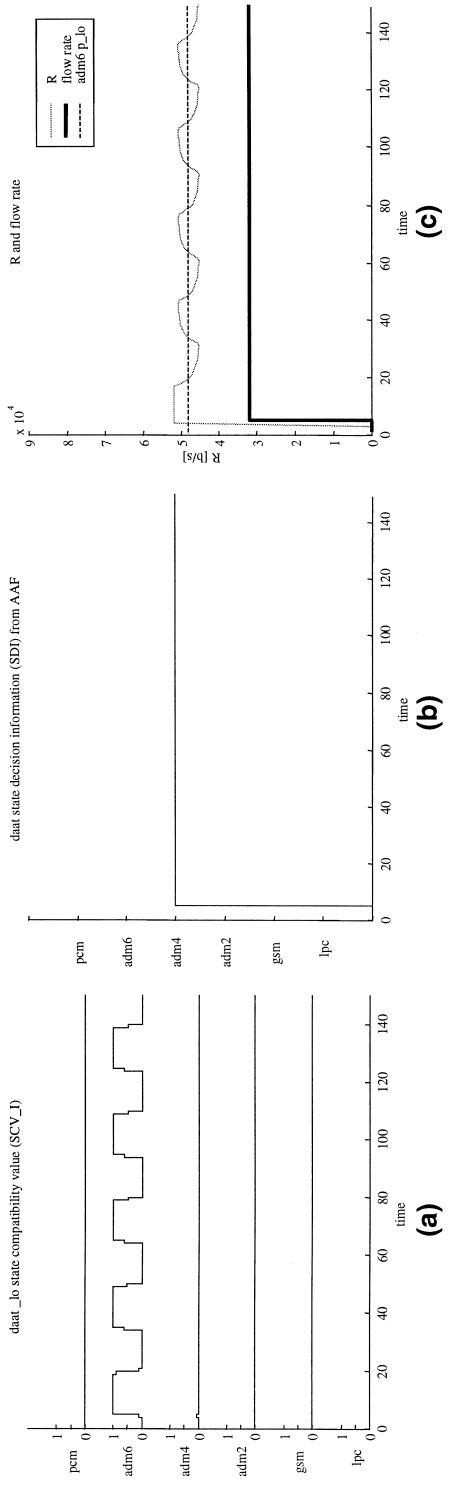

Fig. 14. Using QoSiStates to combat state-flapping.

5.3. Helping to prevent state-flapping

We now use new artificial values of R in order to show how the QoSiStates are used to counter a possible state-flapping scenario. First we generate a scenario where there is state-flapping, as shown Fig. 13 ($q\_time = 1$, $q\_compatibility = 0.8$). Here, QoSiStates are not used. We see in Fig. 13(b) and Fig. 13(c) how the value of R wavers around the *adm6_lo* boundary causing the *daat* to see-saw between *adm4* and *adm6*. (Again, for clarity we have chosen not to simulate noise for R.)

We now chose to employ a policy where if the NetQoSState is such that the QoSState would be operating very close to its boundaries, we chose not to use that QoSState. For the *daat*, this would help to avoid state-flapping when one of the QoSParams in the QoSState was wavering around a *p_lo* boundary of one of the QoSStates. This makes sense for the *daat* as all its QoSStates have boundaries that (effectively) border on each other so it is possible to use a lower quality QoSState to avoid state-flapping. (In Fig. 14, we see how the use of QoSiStates removes state-flapping.) Fig. 13(a) shows the values of the SCVs, which are identical in both cases while, while Fig. 14(a) shows the additional *_lo* QoSiState SCV_I values used only in the scenario of Fig. 14.

The QoSiStates are effectively indicators that a QoSState is operating very close to one of its boundaries, so they may be interpreted as indicators of ''negative compatibility''. (However their exact use and interpretation will be application specific.) We show the use of the same *q_compatibility* and *q_time* values for the *Q_ISCORE* as that for *Q_SCORE*, but different values could be used. Setting lower values for the *q_compatibility* and *q_time* for *Q_ISCORE* would make the AAF more sensitive to operation of a QoSState close to its boundary. This difference between thresholds and timescales between *Q_SCORE* and *Q_ISCORE* could, for example, be used to allow fast detection of operation close to QoSState boundaries, while still maintain flow stability when operation is well within QoSState boundaries. Similar application-level smoothing may be achievable by using a large enough value for *q_time*, but this would then result in lack of responsiveness in adaptability.

6. Discussion

The AAF is the key to the operation of the application. It is the gateway for interaction between the user and the adaptation process of the application. The AAF presented in this paper is mainly to show the use of the SCV/SCV_I information and demonstrate the usefulness of the QoSSpace and QoSState abstractions.

6.1. The QoSStates and QoSSpace

The use of QoSStates simply extends the general information model currently used in methods of describing flow requirements. Flow requirements are often described with performance parameters that must be met in order for the flow to be functional. This is typical for description of resource reservation requirements. However, our treatment assumes that the dynamics, semantics and relationships of QoSStates (inter-flow and intra-flow) are known only to the application. The QoSSpace needs very little semantic knowledge of the QoSStates, and the main requirement is for ''low'' and ''high'' to having meaning for QoSParams.

The QoSSpace treats all QoSParams as orthogonal. Each QoSParam is evaluated individually, in the PCVF, and only then is the SCVF for each QoSState evaluated. So, effectively, the treatment of any one of the QoSParams is identical to the treatment of just a single QoSParam. Although in some cases there may be correlation between QoSParams, this is for the application designer to resolve and define QoSStates appropriately, if required. The general model of the QoSState allows the QoSSpace to be applied in a more diverse manner than just in the use of ''traditional'' QoS parameters. For example, parameters such as battery life, host load and cost could also be used.

The QoSSpace is presented as an abstraction that requires only two pieces of information per QoSParam per flow: an estimate of the current value, p_p, and an estimate of the current variability in that value, v_p. Our simple definition of v_p means that the QoSEngine needs to hold very little historic information for a flow. Additionally, the QoSEngine is separated from the mechanism that is used to provide the values of p_p and v_p. This means that the implementation of the QoSSpace is not con-

strained. It could, for example, be tightly coupled with the application (embedded), implemented as a kernel module or daemon on a host, or implemented as part of a distributed system using middleware. Also, the nature of this model is such that even if dynamic adaptation is not goal, the use of the QoSStates and QoSSpace may be useful in order to detect per-flow QoS violations from QoS parameter measurements.

6.2. Interaction with the user

The QoS assessment capability offered by the QoSSpace allows *daat* to adapt to fluctuating network QoS. The decision is made by the AAF and the *daat* application *automatically*, but includes (static) user preferences. Another application may be more interactive, letting the user make the decision *manually* but present the user with a list of options based on the *Q_SCORE/Q_ISCORE* values or SCVs, allowing the user to make an informed decision.

The AAF is a simple algorithm. The main control issue is the interpretation of the user's wishes, via the user preferences. We have already seen that the QoSSpace has very little semantic knowledge of the flows. Notice that the AAF also has very simple semantic knowledge of the flows in *daat*. The QoS mapping from the user is simple; better quality QoSStates have higher numbers than lower quality QoSStates. The suitability of use of any particular QoSState is evaluated with *Q_SCORES* and *Q_ISCORES* for which ''high'' and ''low'' also have meaning. These are the only semantics that the AAF is aware of, making it a simple and easily implemented algorithm. Such simple QoS mapping between user, application and network, coupled with the simple nature of SCVs makes for easy decision-making and easy programmability in real applications. The main aim of the *daat* examples is to show that the SCVs provided in the QoSReport ease the decision-making process. If simple relationships can be found between user preferences and application-modes and QoSRegions (as in our *daat* examples), the AAF has quite a simple task to perform.

The exact use of the adaptation capability will ultimately depend on the user and the application. We have modelled the *daat* AAF to automatically adjust its flow rate by changing the flow encoding and this is reasonable because studies show that such behaviour in an audio flow does not adversely affect users' perception of quality [24]. This may not be true for all media types and for all users, even if the media is scaleable (e.g. video). In our examples, we have let the user choose *how* and *when* adaptation occurs, mapping values from the user directly to *q_time* and *q_compatibility*. However, this does not preclude these values from being determined by the application through a different interaction with the user.

6.3. The responsibilities of the AAF

The QoSSpace does not attempt to deal with distributed application issues. Decision-making algorithms in a distributed environment could be centralised or distributed. We have shown a simple AAF algorithm that bases decisions about adaptability on information seen by a single *daat* instance, i.e. local information. If *daat* were used in a conferencing scenario, there may be a need to build in application-level signalling into the AAF. However, there is much heterogeneity (network QoS and user preferences) in multicast scenarios. So, even if the adaptation decisions are not made on a local, per-instance basis, there will need to be feedback of local information from the application sites to any centralised/distributed decision-making mechanism. This may have effects on the value of *q_time* (to account for time required for application-level signalling), and so the application may also wish to have some control over its value. So, both *q_time* and *q_compatibility* could be totally under application control. However, it could also be argued that allowing user dissatisfaction to be expressed as an input to the application (to adjust *q_time* and *q_compatbility* values) could also result in a similar effect but under user control [21].

In a multicast scenario, the decision-making process may make use of localised mechanisms, allowing closely located receivers to make adaptation decisions by exchanging QoSReports. Such self-organised, receiver-driven schemes are currently of great importance for scaleable Internet multicast [25,26], and one key element of their success is being able to share information about the QoS that the application instances experience.

We would expect that application specific adaptation functions would be developed as required. The *daat* simulations indicate it is possible for a single flow to adapt to changes in the QoS available to the flow. Where multiple flows share resources with other flows in a best-effort network, we must be conscious of how the flows affect the network and the application behaviour. We must emphasise that the QoSSpace is a system for providing *QoS information summaries* to aid the process of making adaptation decisions. Ultimately, the AAF must make the adaptation *decision* by selecting a QoSState and is responsible for *implementing* the adaptation. So, we would expect the AAF to include mechanisms for ensuring:

- **Fairness**: with respect to the resources available to other flows.
- **Stability**: in considering how the adaptation policy affects the network.
- **Scalability**: in order to allow the decision making process to be amenable for use in distributed applications.

A multicast application might achieve fair-share and congestion control by use of schemes such as [27] that allow multicast traffic to share capacity fairly with TCP traffic. Again, such mechanisms need information about flow and network compatibility, which may be provided by the QoSSpace. Such mechanisms would also need to cater for congestion due to synchronisation effects in QoSRegion changes (e.g. the ''9.00 am effect'') by use of heuristic mechanisms such as slow-start. In our *daat* AAF, this might be by insisting that all *daat* instances must use the *lpc* QoSRegion while $n < q\_time$.

7. Summary and further work

We have shown how it is possible to allow applications to make adaptation decisions automatically and dynamically from measured QoS parameter values. We have considered:

- **The QoSSpace**: a model of the network that allows application flow-states and network QoS to interact. The QoSSpace can produce QoSReports that give an assessment of the relative compatibility of the network QoS and the application's flow-states. The QoSSpace is intended to for general-purpose use. The reports it produces contain **state compatibility values (SCVs)** which are scaleable, simple in nature and easy to understand and manipulate. **QoSStates** are descriptions of the application's flow requirements using **QoSParam** boundaries, providing a simple and adequate QoS mapping between application and network. The QoSSpace does not unduly constrain the design, construction or operation of the application. QoSParams can be used to model quantities other than network QoS parameters, such as batterry power, host load, cost, etc.
- **That dynamic adaptation is possible**: we have used simulations of a dynamically adaptable audio tool (*daat*) to show how the SCVs from the QoSSpace can be used. A simple algorithm, the application adaptation function (AAF) for the *daat*, was used to incorporate user preferences and application requirements into an automatic adaptation policy that controls the operation of the *daat* application. The AAF uses a SCV threshold, *q_compatibility*, and a stability time, *q_time*, to produce an evaluation, the *Q_SCORE/Q_ISCORE*, of the suitability of a QoSState for use by the application. We have showed how it is possible for the *daat* to adapt the rate of its audio flow by changing audio encoding.
- The QoSSpace is designed to be easily integrated into applications. The QoSStates specify the requirements of a flow but the definition of a flow is not constrained and remains application-specific.

The work on the QoSSpace is ongoing. We need to investigate further:

- **Integrated QoS architecture:** we would like to investigate how dynamic adaptability fits into an integrated QoS architecture with other Internet QoS mechanisms, e.g. resource reservation schemes and differentiated services.
- **Interaction with the user**: as we have noted in this dissertation, consideration of the user preferences and requirements is important. We need to investigate how the use of dynamic adaptability changes the way the user and the application must interact. Is it possible to have totally *automatic* dynamically adaptable applications or must there be some interaction with the user during the execution of an application, rather than just the

use of static user preferences? This would involve an investigation of the interface I_{aq} that we have not considered in detail in this work.

- **Distributed applications**: the following issues need to be addressed:
  - When an application *does* have the ability to make adaptation decisions, what are the interactions between instances that affect application mode changes and flow-requirement changes?
  - How does the local interaction with the user affect the decision-making process in distributed instances?
  - What is the interaction of the architectural components to support receiver heterogeneity in the decision making-process? (For example, how might the decision-making process interact with the filtering mechanisms in the application level gateways or in the network?)

Acknowledgements

This work has been funded through the projects QEDANA (HP Laboratories Bristol, UK) and IB-CoBN (CEC ACTS Project AC101).

References

[1] V. Paxson, End-to-end Internet packet dynamics, Proc. ACM SIGCOMM'97, September 1997, pp. 139–152. ftp://ftp.ee.lbl.gov/papers/vp-pkt-dyn-sigcomm97.ps.Z.

[2] V. Paxson, End-to-end routing behavior in the Internet, IEEE/ACM Transactions on Networking 5 (5) (1997) 601–615, ftp://ftp.ee.lbl.gov/papers/vp-routing-TON.ps.Z.

[3] R.H. Katz, Adaptation and mobility in wireless information systems, IEEE Personal Communications 1 (1) (1994) .

[4] L. Delgrossi, C. Halstrick, D. Hehmann, R.G. Herrtwich, O. Krone, J. Sandvoss, C. Vogt, Media scaling for audiovisual communication with the Heidelberg transport system, Proc. ACM Multimedia, June 1993, pp. 99–104.

[5] N. Yeadon, F. Garcia, D. Hutchison, D. Shepherd, Filters: QoS support mechanisms for multipeer communications, IEEE Journal of Selected Areas in Communication 14 (7) (1996) 1245–1262.

[6] V. Hardman, M.A. Sasse, M. Handley, A. Watson, Reliable audio for use over the Internet, Proc. INET'95, Hawaii, USA, 27–30 June 1995, pp. 171–178. http://www.cs.ucl.ac.uk/staff/A.Sasse/inet95audio.ps.

[7] E. Amir, S. McCanne, M. Vetterli, A layered DCT coder for Internet video, Proc. ICIP'96, Lausanne, Switzerland, September 1996. http.cs.berkeley.edu/ ~ elan/pubs/papers/ldct.ps.

[8] H. Schulzrinne, RTP Profile for Audio and Video Conferences with Minimal Control, RFC1890, January 1996.

[9] J. Case, K. McCloghrie, M. Rose, S. Waldbusser, Introduction to Community-based SNMPv2, RFC1901, January 1996.

[10] R. Yavatkar, K. Lakshman, A CPU scheduling algorithm for continuous media applications, Proc. 5th Int. Workshop on Network and Operating System Support for Digital Audio and Video, Durham, NH, April 18–21, 1995, pp. 223–226. http://hulk.bu.edu/nossdav95/papers/yavatkar.ps.

[11] K. Fall, J. Pasquale, S. McCanne, Workstation video playback performance with competitive process load, Proc. 5th Int. Workshop on Network and Operating System Support for Digital Audio and Video, Durham, NH, April 18–21, 1995, pp. 179–182. http://hulk.bu.edu/nossdav95/papers/KevinFall.ps.

[12] A. Campbell, C. Auurecoechea, L. Hauw, A review of QoS architectures, Proc. 4th IFIP Int. Workshop on Quality of Service (IWQS'96), Paris, France, March 1996.ftp://ftp.ctr.columbia.edu/CTR-Research/comet/public/papers/96/CAM96a.ps.gz.

[13] A.T. Campbell, Mobiware: QoS-aware middleware for mobile multimedia networking, Proc. IFIP 7th Int. Conf. on High Performance Networking, White Plains, New York, April 1997. http://comet.ctr.columbia.edu/ ~ mobiware/wireless/PUB/hpn97.ps.Z.

[14] L. Zhang, S. Berson, S. Herzog, S. Jamin, in: R. Braden (Ed.), Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification, RFC2205, September 1997.

[15] L. Wolf, C. Gridwodz, R. Steinmetz, Multimedia communication, Proceedings of the IEEE 85 (12) (1997) 1915–1933.

[16] A. Campbell, G. Coluson, D. Hutchison, Supporting adaptive flows in quality of service architecture, ACM Multimedia Systems Journal (May 1996). ftp://ftp.ctr.columbia.edu /CTR-Research /comet /public /papers/96/CAM96c.ps.gz.

[17] S. Lu, K.-W. Lee, V. Bharghavan, in: A. Campbell, K. Nahrstedt (Eds.), Adaptive Service in Mobile Computing Environments, in Building QoS into Distributed Systems, Chapman and Hall, London, 1997, pp. 25–36.

[18] K. Nichols, S. Blake (Eds.), Differentiated Services Operational Model and Definitions, IETF DIFFSERV WG, work-in-progress, February 1998.

[19] S. Shenkar, D. Clark, D. Estrin, S. Herzog, Pricing in computer network: reshaping the research agenda, ACM Computer Communications Review 26 (2) (1996) 19–43, ftp.parc.xerox.com/pub/net-research/picn.ps.

[20] A. Campbell, G. Coulson, D. Hutchison, A quality of service architecture, ACM Computer Communications Review 24 (2) (1994) 6–27, ftp://ftp.comp.lancs.ac.uk/pub/mpg/MPG-94-08.ps.Z.

[21] B. Landfeldt, A. Seneviratne, C. Diot, User services assistant: an end-to-end reactive QoS architecture, Proc. 6th Int. Workshop on Quality of Service (IWQoS'98), Napa, CA, USA, 18–20 May 1998.

[22] J.-C. Bolot, A. Vega-Garcia, Control mechanisms for packet video in the Internet, Proc. IEEE INFOCOM'96, San Francisco, CA, USA, March 1996, pp. 232–239. http://www.cs.columbia.edu/ ~ hgs/papers/Bolo9603_Control.ps.gz.
[23] V. Hardman, A. Sasse, I. Kouvelas, Successful multi-party audio communication over the Internet, Communications of the ACM, to appear.
[24] J. Hughes, M.-A. Sasse, Internet multimedia conferencing – results from the ReLaTe project, Proc. ICDE World Conf., Pennsylvania State University, 2–6 June 1997. http://www.cs.ucl.ac.uk/staff/A.Sasse/icde.ps.
[25] I. Kouvelas, V. Hardman, J. Crowcroft, Network adaptive continuous-media applications through self organised transcoding, Proc. 8th Int. Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'98), Cambridge, UK, 8–10 July 1998.
[26] S. McCanne, V. Jacobson, M. Vetterli, Receiver driven layered multicast, ACM SIGCOMM'96, CA, USA, August 1996, pp. 117–130.
[27] L. Vicisano, L. Rizzo, J. Crowcroft, TCP-like congestion control for layered multicast data transfer, Proc. IEEE INFOCOM'98, San Francisco, USA, 29 March–2 April 1998.

Saleem N. Bhatti holds a B.Eng.(Hons) in Electrical and Electronic Engineering, a M.Sc. in Data Communication Networks and Distributed systems, and a Ph.D. in Computer Science, all from University College London (UCL). Since October 1991 he as been a member of the research staff at UCL, and has recently been appointed as Lecturer in Data Communications and Networking within the department. He has been involved in post-graduate teaching as well as various consultancy roles within industry in the area of networking related projects, covering topics such as network and systems management, network security, ISDN, ATM, QoS, multi-service networks and residential broadband services. Saleem is a member of the IEEE and the ACM.

Graham Knight graduated in Mathematics from the University of Southampton in 1969. He taught Maths and Computer Science in schools and colleges until 1979 when he obtained a M.Sc. in Computer Science from University College London. Since then, he has worked at the college, initially as a Research Assistant, now as a Senior Lecturer. His research work has been in the field of computer communications. He has led UCL teams in ESPRIT projects concerning OSI, network management and ISDN, and is now involved in European ACTS programme. Currently his research interests include routing in hybrid SDH/ATM networks, CATV-based Internet access and QoS, and the use of Java and WWW for network and systems management.